icfes saber pro

INFORMES

Prueba de Habilidades Genéricas **GSA Colombia**

Resultados del pilotaje

Presidente de la República
Juan Manuel Santos Calderón

Ministra de Educación Nacional
María Fernanda Campo

Viceministro de Educación Superior
Javier Botero Álvarez

Directora General
Margarita Peña Borrero

Secretaria General
Gioconda Piña Elles

Jefe de la Oficina Asesora de Comunicaciones y Mercadeo
Ana María Uribe González

Director de Evaluación
Julián Patricio Mariño Von Hildebrand

Director de Producción y Operaciones
Francisco Ernesto Reyes Jiménez

Director de Tecnología
Adolfo Serrano Martínez

Subdirectora de Análisis y Divulgación
Maria Isabel Fernandes Cristóvão

Elaboración del documento
Ignacio Gómez Montes

Revisión técnica
Maria Isabel Fernandes Cristóvão

Revisión de estilo
Felipe Solano Fitzgerald

Diseño
Giovanni Camacho Solorza

ISBN de la versión electrónica: 978-958-11-0549-6

Bogotá, D.C., enero de 2011

Tabla de contenidos

Introducción

La prueba de habilidades genéricas (Graduate Skills Assessment, GSA) fue diseñada por el Consejo Australiano para la Investigación Educativa (Australian Council for Education Research, ACER), uno de los centros de investigación y evaluación educativa líder en el mundo. El ACER ha participado en aplicaciones de esta prueba dentro y fuera de Australia, en diferentes proyectos en países como Reino Unido, Chile, Cambodia, Indonesia, Emiratos Árabes, entre otros, así como también en el desarrollo de pruebas internacionales comparadas como PISA – Programa Internacional de Evaluación de Estudiantes e ICCS – Estudio Internacional de Educación Cívica y Ciudadana.

A finales de 2008 el Instituto Colombiano para la Evaluación de la Educación (ICFES) contrató con el ACER una licencia de uso de este examen. La GSA consiste en cuatro pruebas creadas para medir un conjunto de habilidades genéricas[1] ampliamente aplicables que pueden ser desarrolladas por los estudiantes universitarios. Esta evalúa las dimensiones de resolución de problemas, pensamiento crítico, comunicación escrita y entendimiento interpersonal. Las pruebas correspondientes a estas dimensiones se estructuraron a partir de encuestas realizadas a académicos y empresarios australianos sobre los requisitos necesarios para continuar estudios superiores o ingresar al mercado laboral.

Las distintas dimensiones de la GSA tienen como propósito identificar y caracterizar las habilidades genéricas de los estudiantes universitarios, las cuales resultan clave tanto para la formación académica como para el futuro desempeño laboral. El hecho de que el resultado en una prueba corta de habilidades genéricas como la GSA esté correlacionado positivamente con medidas de éxito en la universidad y en el trabajo profesional –lo cual a su vez está asociado con una amplia gama de conocimientos y habilidades curriculares– otorga gran importancia a tales habilidades en la validación del desempeño académico; además, aporta a la validez de la prueba y su aplicación en Colombia por parte del ICFES.

[1] La Australian Technological Network of Universities (ATN), que comprende cinco universidades australianas (Curtin University of Technology, University of South Australia, RMIT University, University of Technology Sydney y Queensland University of Technology), define las competencias genéricas con el término *graduate attributes* ('atributos de los graduados') como: "las cualidades, habilidades y comprensiones que una comunidad universitaria acuerda que sus estudiantes deben desarrollar durante el tiempo en el que estén en la universidad. Estos atributos incluyen la destreza o conocimiento técnico de la disciplina que tradicionalmente ha formado parte del núcleo de los cursos universitarios, pero van más allá. Son cualidades que preparan a los graduados como agentes del bien social en un futuro desconocido". Véase Hambur, Sam, et. al. (2002). *Graduate Skills Assessment Stage One Validity Study*, Commonwealth Department of Education, Science and Training, Australian Council for Educational Research, Commonwealth of Australia.

En junio de 2008, la Dra. Jennifer Bryce, especialista en evaluación del ACER, presentó la GSA a la comunidad académica colombiana. El ICFES decidió entonces realizar un pilotaje de la prueba en el país. Para ello, el 26 de febrero de 2009 la aplicó a una muestra de estudiantes de primer y último año de diferentes programas académicos de 16 instituciones de educación superior. La asesoría del ACER en el diseño y la implementación del pilotaje permitió evaluar el comportamiento de los distintos componentes de la GSA y dio bases para avanzar en la definición y diseño de un examen de egreso común para todos los núcleos básicos de la educación superior en Colombia.

Tanto la traducción de la prueba al español como toda la logística asociada con su aplicación estuvo a cargo del ICFES. La traducción fue revisada por el ACER, entidad que también proporcionó la versión en inglés del *Manual de Supervisores*; el *OMR* de GSA para el registro de las respuestas de selección múltiple, la *Cartilla de Comunicación Escrita* para las respuestas escritas y el *Reporte de Prueba de GSA*. Por su parte, el ICFES realizó la administración de la prueba, mientras que el ACER proporcionó a un miembro de su equipo para adelantar actividades de capacitación de tres días de duración para marcadores (*marker training*)[2]. Finalmente, como parte de los esfuerzos conjuntos entre el ACER y la Dirección de Evaluación del ICFES, se llevó a cabo un taller de calificación del componente de comunicación escrita de la prueba del 23 al 25 de marzo de 2009, actividad que se orientó al grupo de calificadores de la aplicación del piloto.

Este informe recoge los resultados del proceso de aplicación del piloto de la prueba GSA en Colombia y está dividido en cuatro capítulos. En el primero se presentan las características generales de la prueba en cuanto a sus objetivos, usos, formatos, estructura, así como la descripción de las dimensiones evaluadas, el método de evaluación y la relevancia y la validez de la misma. En el segundo se describe el marco muestral; en el tercero se presentan los resultados obtenidos en cuanto a desempeños de los estudiantes y análisis de los ítems y en el último se plantean las principales conclusiones derivadas del pilotaje.

[2] En marzo de 2009 el ACER condujo durante tres días un taller de entrenamiento dirigido a marcadores (*marker training*) en Bogotá para el componente de comunicación escrita de la prueba GSA-Colombia. Las guías de marcación, adaptadas de las versiones en inglés usadas en Australia, fueron preparadas tanto para las tareas de reporte como de argumento, y luego fueron traducidas al español. Se capacitó a marcadores y supervisores en las demandas de cada tarea y en las correspondientes guías de marcación. Se seleccionaron unos escritos (*scripts*) de estudiantes en ambas actividades para práctica y discusión. Los escritos (*scripts*) cubrían la totalidad de rangos de desempeños de los estudiantes. Después los marcadores trabajaron individualmente bajo la dirección de un supervisor, cuyo papel era resolver las diferencias y los problemas de marcación. Aproximadamente diez marcadores fueron entrenados en la marcación de cada tarea, junto con tres supervisores. Se dedicó un día a la práctica y discusión de escritos para cada tipo de ejercicio. El tercer día se usó para práctica y solución de diferencias y asuntos que surgieron durante la marcación. Al finalizar el taller se logró un buen nivel de acuerdo entre los marcadores.

1. Características de la prueba GSA

1.1 Usos de la prueba

En la GSA el término “habilidades genéricas” significa la habilidad concreta de los estudiantes para desarrollar efectivamente ciertos tipos de tareas que involucran un razonamiento genérico superior. Las habilidades genéricas son entonces transferibles y aplicables al trabajo de los niveles universitario y de profesional graduado. La transferibilidad significa que las habilidades genéricas se aprenden en un contexto determinado y su estructura varía según cada disciplina. Esto quiere decir que la familiaridad con el contexto del estudiante influye en su motivación y desempeño. Por tanto la prueba se diseñó para ser aplicada en contextos diferentes, suponiendo que si un estudiante puede desarrollar habilidades genéricas en un ámbito familiar, existe mayor probabilidad de que pueda llevar a cabo tareas efectivas en espacios diferentes con los cuales adquiera familiaridad.

La prueba puede ser usada como herramienta de diagnóstico en el momento de ingreso a la universidad y sus resultados permiten, por ejemplo, dar apoyo a aquellos estudiantes que lo necesiten en alguna de las áreas. En cuanto a la culminación de los estudios universitarios, la prueba GSA resulta útil tanto para empleadores en la selección laboral como herramienta de admisión a cursos de posgrado. La prueba también puede ser aplicada por medios computarizados, mediante un banco de ítems que permite, entre otras cosas, asignar componentes seleccionados a estudiantes elegidos (por ejemplo, la prueba puede modificarse para cada área del conocimiento de manera que los estudiantes de determinado programa respondan ítems específicamente enfocados en esa área; de igual manera, los departamentos de las universidades pueden incluir conjuntos de ítems de interés para ellos). Además, la GSA tiene el potencial de servir a las universidades como herramienta para la predicción del desempeño académico de los alumnos para admisiones a cursos de pregrado y posgrado, por ejemplo, lo cual podría hacerse dando cierto peso a los componentes de la prueba para optimizar las predicciones. La prueba también podría ser empleada para dar puntajes equivalentes a notas de la universidad para aquellos estudiantes que no las tengan.

En general la prueba sirve para comparar las habilidades académicas frente al desempeño, lo cual involucra elementos comunes a los distintos contextos en que pueda ser aplicada, a saber: comprensión verbal sociocultural, que involucra el contexto de las humanidades y ciencias sociales; la lógica y el razonamiento cuantitativo, que se usan en contextos de matemáticas aplicadas, ciencias naturales y ciencias sociales, y la comunicación escrita, aplicable a todos los contextos.

1.2 Formato y estructura

La prueba GSA utiliza tanto preguntas de selección múltiple como abiertas o de respuesta construida; ambas tienen ventajas y desventajas. Por una parte, las de selección múltiple permiten escoger opciones de respuesta complejas en las que los estudiantes deben generar soluciones y/o puntos de vista que correspondan con la opción seleccionada para lograr un buen desempeño en la prueba, sin limitarse a adivinar la respuesta. Además, este tipo de preguntas evita la variabilidad entre los marcadores. Por otra parte, las preguntas de respuesta abierta permiten generar y aplicar soluciones y puntos de vista, a partir de niveles mínimos de incitación o instrucciones, tal como sucede en la mayoría de las tareas de la vida real; por lo tanto son útiles en aquellas situaciones en las que es necesario plantear o aplicar soluciones efectivas. Sin embargo, se requiere tener especial cuidado en la selección de los marcadores para evitar variabilidad no debida, riesgo del que quedan libres los ítems de selección múltiple.

La prueba se creó para medir psicométricamente diferentes conjuntos de habilidades (dimensiones) que son relevantes para el trabajo académico y la industria. Para tal fin el ACER diseñó un constructo que se enfoca en cinco dimensiones cognitivas, aceptando que las habilidades genéricas deben tener un alto grado de transferibilidad a contextos distintos una vez se adquiera familiaridad. Las dimensiones cognitivas seleccionadas son: pensamiento crítico, solución de problemas, entendimiento interpersonal, redacción argumentativa y redacción de un reporte. Esas dimensiones cognitivas o componentes se escogieron atendiendo a criterios de popularidad[3], que tuviesen significado psicométrico, que se constituyesen en elementos esenciales de otras habilidades, que fuesen transferibles y fácilmente mensurables. Las primeras tres secciones de la prueba GSA consisten en preguntas de selección múltiple, con 30 ítems por dimensión, que deben ser resueltos en dos horas. La dimensión de comunicación escrita está constituida por dos tareas planteadas mediante preguntas abiertas para ser resueltas en una hora; los estudiantes deben elaborar dos documentos: un reporte y una argumentación.

[3] La "popularidad" se refiere a aquellas habilidades sugeridas por todos los consultados (stakeholders entre los que había universidades, empresas o empleadores, consejeros profesionales, etc.) que tuvieron un mayor número de respuestas durante el proceso de escogencia de los componentes o dimensiones cognitivas en Australia. Estas fueron comunicación escrita, pensamiento crítico, solución de problemas y entendimiento interpersonal porque, al parecer, constituían elementos de otras habilidades (como la capacidad de aprendizaje de largo plazo - "life-long learning") y porque eran fácilmente transferibles y mensurables. Véase Hambur, Sam, et.al, op. cit., 2002.

1.3 Dimensiones de la prueba GSA

1.3.1 Pensamiento crítico

El enfoque del pensamiento crítico se centra en razonar sobre problemas de la vida diaria. Se pide a los estudiantes que comprendan, analicen y evalúen afirmaciones / fragmentos que representan puntos de vista actuales sobre el mundo real. Debido a que la habilidad de pensar críticamente depende de la familiaridad con el contexto, los ítems utilizados en la prueba GSA tienden a ser generalmente accesibles, aunque hay evidencia de lenguaje especializado, material científico o matemático. Para diferenciar el pensamiento crítico de la solución de problemas, que también contiene aspectos del primero, se presentan los puntos de vista en forma de texto, evitando usar material cuantitativo, de tal manera que aquellos aspectos relacionados con interpretación de datos se evalúan en el componente de solución de problemas.

Dado que el componente debe ser psicológicamente coherente, se evita el uso de material no verbal. Los ítems son de selección múltiple y pueden ser categorizados (aunque un ítem puede tener facetas en más de una categoría) como:

- Comprensión para identificar significados explícitos e implícitos.

- Análisis e inferencia para identificar las definiciones a ser aplicadas, reclamos, puntos de vista, asuntos clave, líneas de razonamiento, evidencias, conclusiones, argumentos, supuestos, fallas lógicas, implicaciones lógicas, información faltante, instrumentos retóricos, ambigüedades, analogías, etc.

- Síntesis y evaluación para juzgar aspectos como la credibilidad y validez de la evidencia, líneas de razonamiento, conclusiones y argumentos.

La dimensión de pensamiento crítico de la GSA busca proporcionar información sobre la habilidad para pensar críticamente acerca de puntos de vista sobre distintos asuntos y tomar decisiones basadas en estándares intelectuales internacionalmente aceptados. El pensamiento crítico se centra en los estándares intelectuales y en los elementos de pensamiento y razonamiento que permiten comprender, analizar y evaluar los puntos de vista presentados en el texto. El desempeño en pensamiento crítico corresponde a la habilidad para organizar y controlar el pensamiento efectiva y sutilmente, incluyendo la habilidad para generar criterios de evaluación adecuados.

1.3.2 Entendimiento interpersonal

Esta es un área compleja y en evolución. El componente de entendimiento interpersonal de la prueba GSA trata solamente un aspecto limitado de este campo. Los ítems son presentados usualmente como textos (generalmente bajos en cuanto a la demanda de habilidades verbales), aunque también puede ser usado material gráfico. Los ítems son de selección múltiple y se enfocan en la habilidad de los estudiantes para:

- Discernir los sentimientos, motivaciones y el comportamiento de otras personas, así como de otros asuntos relacionados con ayudar o trabajar con otros.

- Reconocer cómo puede ser aplicado ese discernimiento de manera tal que efectivamente se pueda ayudar o trabajar con otros, incluyendo retroalimentación efectiva, atención, comunicación, negociación, trabajo en equipo y liderazgo.

La prueba GSA no está interesada en medir la inteligencia per se, sino principalmente en dar información sobre las habilidades que pueden ser deliberadamente desarrolladas en la experiencia universitaria, incluyendo habilidades interpersonales. Acorde con este propósito, la dimensión de entendimiento interpersonal busca medir la destreza de los individuos en aspectos de relaciones interpersonales relevantes para trabajar y vivir efectivamente en comunidad.

Aquellos estudiantes con un entendimiento amplio y sofisticado de los asuntos interpersonales, especialmente aquellos relacionados con el trabajo en equipo y habilidades metacognitivas con enfoque social, tendrán mayores probabilidades de lograr un buen desempeño en la prueba.

1.3.3 Solución de problemas

La prueba GSA se ha enfocado en problemas de la vida diaria con complejidad variable, así como en la habilidad de los estudiantes para identificar, analizar, interpretar, traducir, reorganizar y aplicar apropiadamente la información relacionada con ellos. Se espera que los evaluados utilicen un enfoque lógico y organizado en el análisis y aplicación de la información relevante.

Se presume un mínimo nivel de matemáticas –aunque no se tratan problemas de matemática especializada–, de habilidad interpersonal o de administración de negocios. Los ítems son de selección múltiple y se requiere que los estudiantes:

- Identifiquen, comprendan y reformulen el problema.
- Identifiquen y analicen la información relevante para el problema.
- Representen características del problema.
- Traduzcan, reorganicen, sinteticen y apliquen información relevante para el problema.
- Conceptualicen / generen estrategias y sus resultados.

La dimensión de solución de problemas de la prueba GSA mide la habilidad del estudiante para analizar y transformar la información de manera que le permita progresar en la solución de problemas. Ahora bien, en términos psicométricos es factible que la forma como se presenta la información y el tipo de problema, así como su contenido específico, sean relevantes para su solución por parte de una persona con un conocimiento particular. Se espera que los estudiantes demuestren un enfoque lógico y organizado en el análisis y la aplicación adecuada de información, por lo cual se requiere la puesta en práctica de procesos de razonamientos analíticos, lógicos, cuantitativos y metaestratégicos.

Si bien se presume un conocimiento básico numérico, incluyendo algoritmos aritméticos simples, todos los problemas de la prueba pueden resolverse con conocimientos matemáticos primarios y se permite el uso de calculadoras.

1.3.4 Comunicación escrita

El diseño de las tareas de escritura se enfocó en dos géneros (argumento y reporte) que se valoran según el rango de facultades y lugares de trabajo, y que son apropiados para medir la madurez y experiencia de los estudiantes de educación superior. Estas formas de escritura se consideran un formato adecuado para la prueba GSA porque se basan en aspectos de utilización de habilidades genéricas, en particular obtener, analizar y organizar información, comunicar ideas e información y planificar.

El reporte es una forma común de comunicación escrita sustantiva en la universidad y la vida profesional. Las habilidades para elaborar este tipo de documento se requieren comúnmente y son ampliamente aplicables. La tarea del reporte requiere que los estudiantes comprendan, seleccionen, organicen y presenten claramente información fáctica.

La escritura del argumento se menciona comúnmente en las descripciones de habilidades esperadas de los graduados. La habilidad para presentar un argumento de manera clara,

concisa, lógica y con sentido es solicitada comúnmente y ampliamente aplicable. Esta tarea requiere que los estudiantes desarrollen su punto de vista sobre algún asunto y que estructuren y presenten un argumento que lo apoye.

El material de estímulo para cada tarea proporciona a los estudiantes una plataforma de información o datos para construir su escrito. También se dan instrucciones claramente expresadas y diferenciadas acerca del ejercicio y los criterios con los que se juzgará el texto. El estímulo del argumento es un conjunto de opiniones o comentarios relacionados con un asunto social, mientras que el del reporte consiste en datos presentados en gráficos y tablas.

La dimensión de comunicación escrita de la GSA se desarrolló, entonces, para elaborar descripciones de lo que se espera de los estudiantes. Las formas de escritura fueron escogidas según aspectos de habilidades genéricas relacionadas con trabajo académico y profesional, particularmente recolectar, organizar, analizar y comunicar información, además de planificar. Aquellos estudiantes con habilidades en comunicación escrita, incluyendo aquellas asociadas con el análisis, organización y presentación clara de la información y los puntos de vista, así como con habilidades metacognitivas bien desarrolladas que permitan una aplicación efectiva de las anteriores, tienen mayores probabilidades de desempeñarse bien en la prueba.

1.4 Método de evaluación

En Australia, la prueba GSA usa dos métodos para indicar el desempeño de los estudiantes, que se reporta en formatos diseñados para tal efecto. El primero establece rangos de referencia para comparar el rendimiento entre estudiantes en un campo particular del conocimiento. Para esto proporciona el 60% de todos los puntajes de estudiantes y lo compara con el 60% de los puntajes de estudiantes en campos similares del conocimiento. El segundo método entrega descriptores de niveles de desempeño y puntajes.

La prueba GSA que se piloteó en Colombia tiene, en general, tres niveles de desempeño, para los cuales hay identificados unos descriptores. Estos niveles se determinaron tanto sobre la base de juicios preliminares de las habilidades apropiadas, como sobre el desempeño de los estudiantes. El **nivel 1** (entre 200 y 325 puntos) sugiere que un estudiante tiene unas habilidades básicas o limitadas frente a alguna de las dimensiones particulares medidas en las pruebas. El desempeño en el **nivel 2** (entre 325 y 475 puntos) indica que un estudiante tiene habilidades relativamente sólidas frente a alguna de las dimensiones particulares, mientras que el desempeño en el **nivel 3** (por encima de 475 puntos) significa que un estudiante demuestra una gran maestría en habilidades relevantes para alguna de las dimensiones de la prueba. Además, los puntajes inferiores a 200 se reportan como del nivel 1, pero se especifica que no alcanzan a llegar a un estándar mínimo o limitado, en tanto que

los puntajes superiores a 600 se reportan como de nivel 3, aclarando que superan esa cifra y, por lo tanto, corresponden a un alto grado de maestría.

En las **Figuras 1 y 2** se enuncian las competencias típicas que muestran las personas ubicadas en cada nivel específico. Se espera que una persona que ha alcanzado determinado nivel demuestre las competencias de ese nivel, así como las de los niveles inferiores.

Figura 1. Descriptores para cada nivel en las dimensiones de pensamiento crítico y entendimiento interpersonal

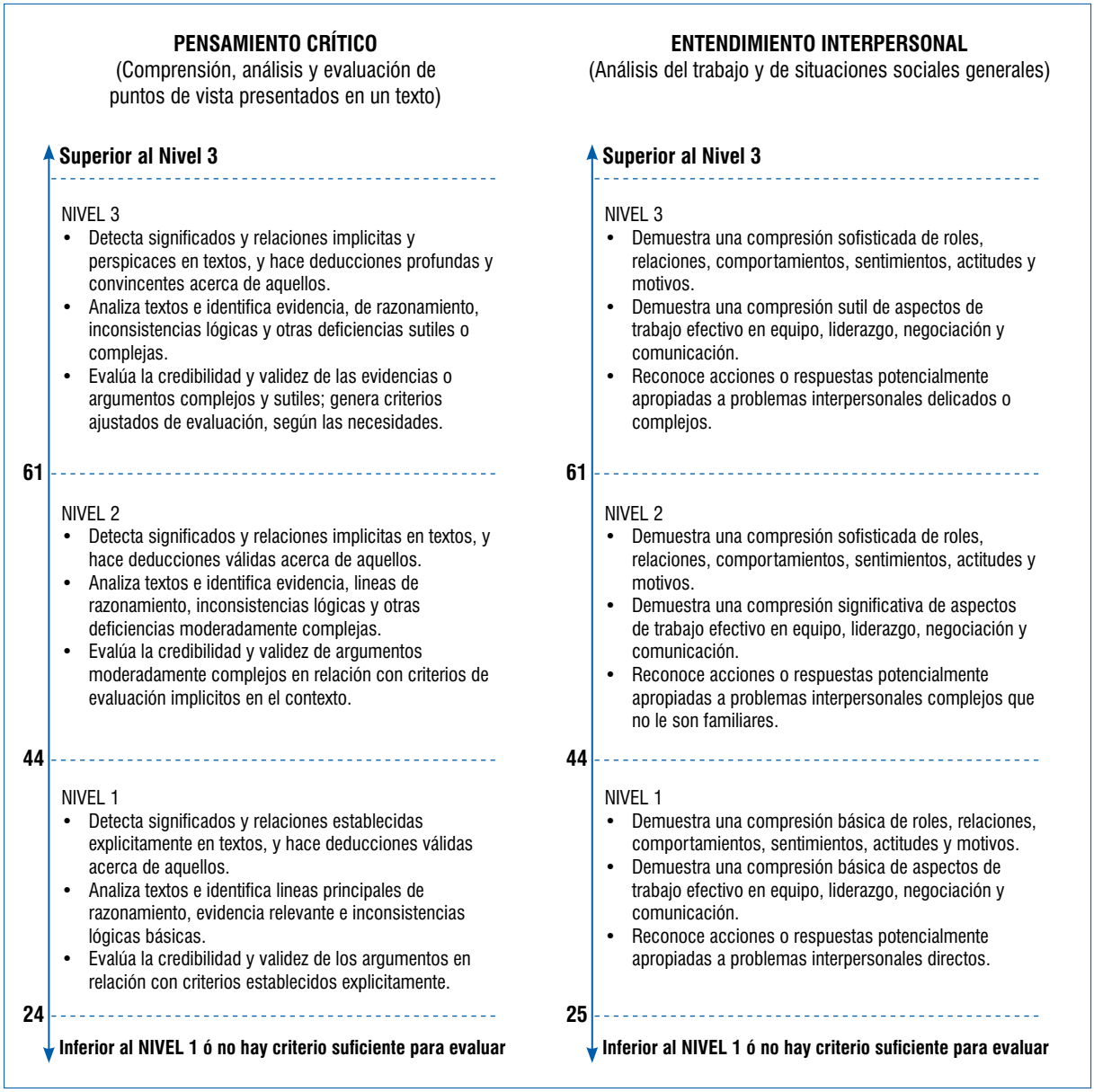
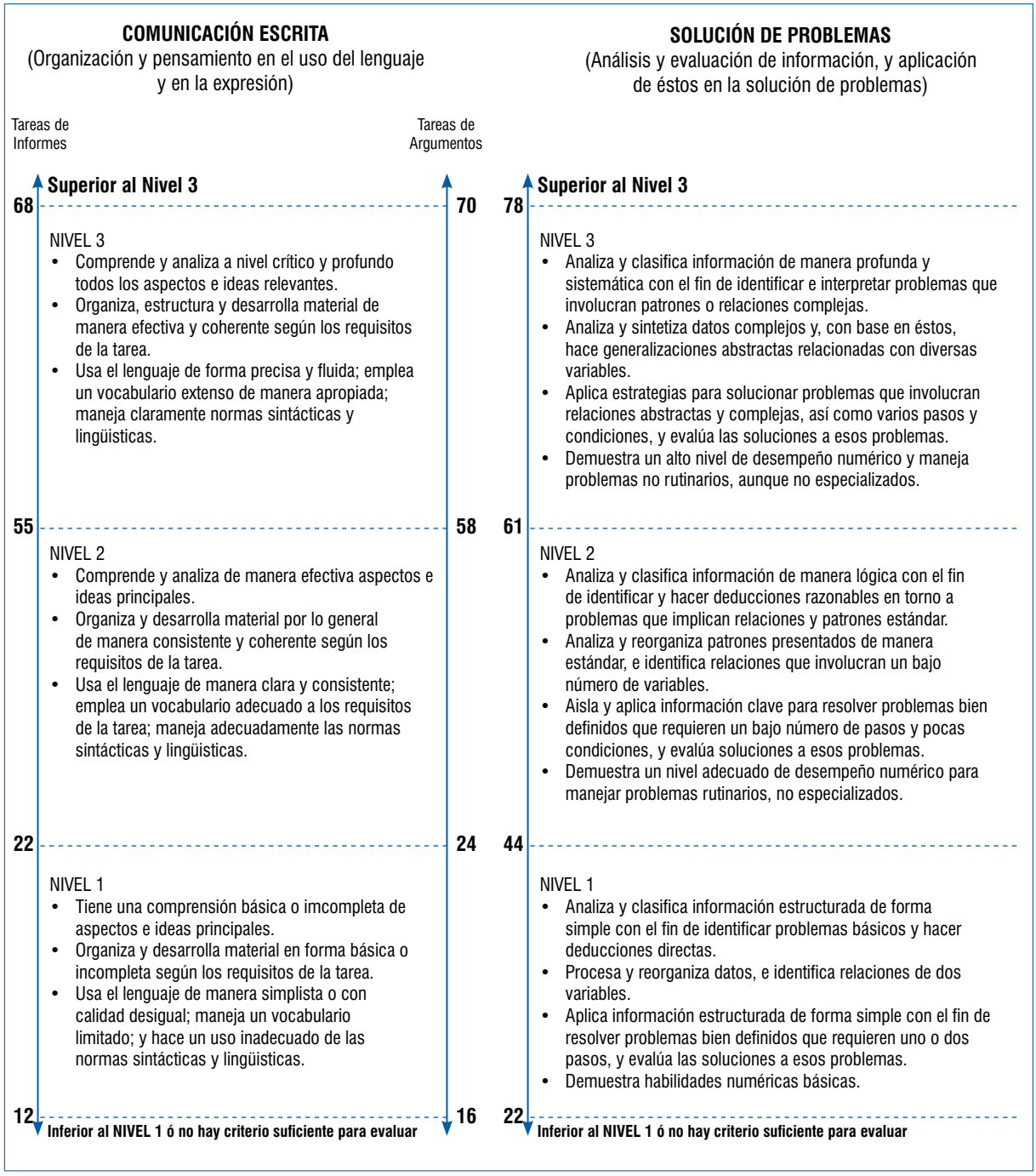


Figura 2. Descriptores para cada nivel en las dimensiones de comunicación escrita y solución de problemas



Aunque los componentes de selección múltiple de la prueba GSA son confiables para muchos propósitos, como medir cambios relativamente pequeños de desempeño entre los grupos de estudiantes que entran a la universidad y los que salen, pueden no ser satisfactorios para valorar pequeños cambios en el desempeño individual de cada estudiante.

La calificación del componente comunicación escrita, por su parte, requirió un entrenamiento especial que el ACER efectuó a un grupo de calificadores de preguntas abiertas.

1.5 Relevancia y validez

La relevancia ha sido estudiada en Australia, donde se ha dado un amplio debate en el que las discusiones en torno a la prueba evidencian una diferencia dramática de posiciones sobre aspectos de ésta que hasta cierto punto se relacionan con el perfil académico detrás de cada enfoque (por ejemplo, enfoques desde las humanidades frente a las ingenierías). En general, las preocupaciones de las distintas formaciones tienen que ver con la posibilidad de sacar tablas de liga; sobre si hay niveles de habilidades genéricas fuera de las disciplinas profesionales; la privacidad de los resultados; la pregunta sobre si las universidades de hecho enseñan esas habilidades; las limitaciones de la selección múltiple; la relevancia de las habilidades interpersonales para investigadores; las especificaciones de las audiencias; las plataformas para la escritura; la relevancia de la prueba para todos los estudiantes universitarios y frente al desempeño profesional, los sesgos culturales, entre otros.

No obstante lo anterior, los argumentos a favor son concluyentes. En el caso australiano, a varios expertos constructores de contenido y a otros interesados, entre ellos estudiantes que habían presentado la prueba, se les pidió que evaluaran la relevancia del contenido y una muestra de ítems. En general, los consultados respondieron que la prueba es positiva en la validez del contenido; sin embargo, plantearon inquietudes frente a la relación entre el desempeño universitario y profesional / laboral y el resultado obtenido en la prueba.

Se entiende que se debe prestar especial atención a refinar la validez de los descriptores de contenido y de nivel. Además, se requiere tener presente la necesidad de asesoría continua y seguimiento a la cuestión de la validez a medida que la prueba se desarrolla; esto es importante ante el reto de producir un examen de habilidades genéricas validado empíricamente que satisfaga a un amplio rango de interesados con demandas substanciales.

Para ser válida, la prueba GSA debe desarrollarse con las siguientes condiciones: la construcción de ítems debe haber sido validada por interesados y expertos; la estructura de la prueba ha de tener una dimensión discriminante y significativa; asimismo, entre las variables, la relacionada con el desempeño de estudiantes debe ser significativa e incluir

información sobre número de años en la universidad en una disciplina particular. El desempeño en la prueba GSA debe estar relacionado con el rendimiento en otras pruebas que miden habilidades similares (como por ejemplo el ECAES, en el caso colombiano) y no debe estar afectado por variables como raza o género. Además, la prueba GSA debe tener rangos de referencia confiables según su propósito.

Sin importar la amplitud del concepto de inteligencia ni los debates y la extensa literatura al respecto, el desempeño en la prueba GSA se ve claramente influenciado por esta, bien sea esté afectada por factores genéticos y/o del ambiente o bien sea que exista una inteligencia ejecutiva y/o inteligencias modulares. Finalmente, además de que intervienen aspectos como la motivación y la confianza, lo importante es que el desempeño en la prueba GSA está afectado por la experiencia universitaria y está relacionado con el rendimiento académico y profesional.

2. Objetivos del pilotaje y marco muestral de la prueba GSA en Colombia

2.1 Objetivos

La prueba GSA tiene los siguientes objetivos establecidos para el caso australiano: investigar la estructura de validez discriminante, identificar variables relacionadas con el rendimiento diferenciado en la prueba, investigar la relación entre el desempeño de los estudiantes y otras medidas de desempeño (otras pruebas), considerar la pertinencia de los rangos de referencia y evaluar la validez del contenido del constructo y los ítems.

El principal objetivo del piloto de la versión colombiana de la prueba GSA consiste en investigar si el instrumento aplicado en el país mantiene su validez discriminatoria para identificar y caracterizar las habilidades genéricas de los estudiantes universitarios; evaluar la validez de esa afirmación y los ítems de la prueba en el contexto nacional, revisar la idoneidad de los niveles de competencia definidos e identificar variables relacionadas con las diferencias de desempeño.

2.2 Marco muestral

El marco muestral del piloto realizado en Colombia se conformó con instituciones de educación superior y programas registrados en el archivo histórico de las pruebas ECAES para las cinco principales ciudades del país. De ese marco muestral se escogieron 20 instituciones para conformar la muestra. También se definió una muestra de programas académicos para cada una de las siguientes cuatro áreas del conocimiento: administración, contaduría y economía; educación; ingeniería y arquitectura; y salud.

El diseño muestral determinó dos requisitos mínimos. Primero, que al menos 3.200 estudiantes de primer y último año presentaran el examen (1.600 alumnos en cada caso) y, segundo, que esos estudiantes estuvieran distribuidos en grupos de 400 que cursaran primer año y 400 de último año, en cada área del conocimiento. El número de personas que presentaron los ECAES en 2008 sirvió como punto de referencia para este diseño. Conforme con el primer requisito, la prueba finalmente se realizó a una muestra de 3.757 estudiantes. Sin embargo, el segundo requisito no se cumplió para los estudiantes de último año para las áreas de administración, contaduría y economía, ni para las ingenierías. Durante el piloto, 5 de los 3.757 estudiantes no realizaron el componente de selección múltiple. La tasa de respuesta de los alumnos que presentaron el examen, frente a los que se citaron, fue del 75%.

Las **Tablas 1a** y **1b** presentan la información sobre el número total de evaluados en cada una de las áreas del conocimiento según el año cursado. La sumatoria total de estudiantes de todas las áreas corresponde a 3.479. La diferencia con respecto a los 3.757 evaluados reportados anteriormente se debe a la exclusión de quienes no reportaron el semestre que estaban cursando y de aquellos matriculados en los semestres tercero, cuarto y quinto.

Tabla 1a. Número total de evaluados por programa en las áreas de administración, contaduría, economía y educación

ÁREA	PROGRAMA	1er. AÑO	ÚLTIMO AÑO
Administración	Administración de Empresas	213	193
Contaduría	Contaduría	132	93
Economía	Economía	75	85
	Total área	**420**	**371**
Educación	Lic. educación básica	47	24
	Lic. educación infantil	151	100
	Lic. educación especial	23	59
	Lic. educación física	60	81
	Lic. en ciencias naturales	0	42
	Lic. en ciencias sociales	34	35
	Lic. en humanidades y lengua castellana	0	56
	Lic. en inglés	3	10
	Lic. en matemáticas y física	19	6
	Lic. en matemáticas	1	17
	Lic. en lenguas modernas	72	71
	Total área	**410**	**501**

Tabla 1b. Número total de evaluados por programa en las áreas de ingeniería y salud

ÁREA	PROGRAMA	1er. AÑO	ÚLTIMO AÑO
Ingenierías	Ingeniería ambiental	48	31
	Ingeniería civil	147	46
	Ingeniería de sistemas	122	89
	Ingeniería eléctrica	17	6
	Ingeniería electrónica	4	17
	Ingeniería industrial	67	130
	Ingeniería mecánica	76	11
	Total Área	**481**	**330**
Medicina	Medicina	130	112
Otros salud	Bacteriología	28	41
	Enfermería	98	127
	Fisioterapia	75	148
	Fonoaudiología	0	16
	Odontología	117	34
	Optometría	20	20
	Total Área	**468**	**498**

La **Tabla 2** muestra la distribución de instituciones y programas seleccionados por ciudad, y la **Tabla 3** permite observar la cantidad de programas que conforman el marco muestral, según área de estudio, así como los seleccionados en la muestra.

Tabla 2. Número de instituciones y programas seleccionados por ciudad

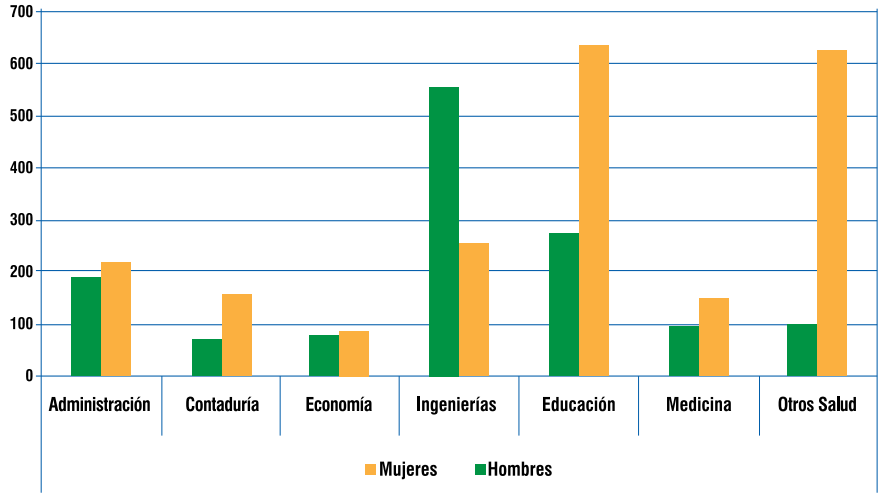
MUNICIPIOS	CANTIDAD DE INSTITUCIONES	CANTIDAD DE PROGRAMAS
Barranquilla	7	64
Bogotá	50	318
Bucaramanga	8	58
Cali	10	76
Medellín	21	119
Total	**96**	**635**

Tabla 3. Número de programas en el universo y en la muestra

ÁREA	PROGRAMAS EN MUESTRA	PROGRAMAS EN EL UNIVERSO	PORCENTAJE DE MUESTRA DE PROGRAMAS
Administración, contaduría y economía	42	202	20,8
Ciencias de la salud	15	89	16,9
Educación	30	84	35,7
Ingeniería y arquitectura	39	260	15,0
Total general	**126**	**635**	**19,8**

En cuanto a la distribución de la muestra por género, se observa en general que el número de mujeres supera la cantidad de hombres que presentaron el examen, especialmente en las áreas de educación y salud diferentes a medicina. En el área de ingeniería el número de hombres fue mayor. En total 60% de las personas evaluadas fueron mujeres y el 40% hombres. En el **Gráfico 1** se presenta la población evaluada según su género y el programa cursado.

Gráfico 1. Población evaluada en el piloto de la prueba GSA según género y programa



La **Tabla 4** muestra la distribución de los evaluados por institución y programa.

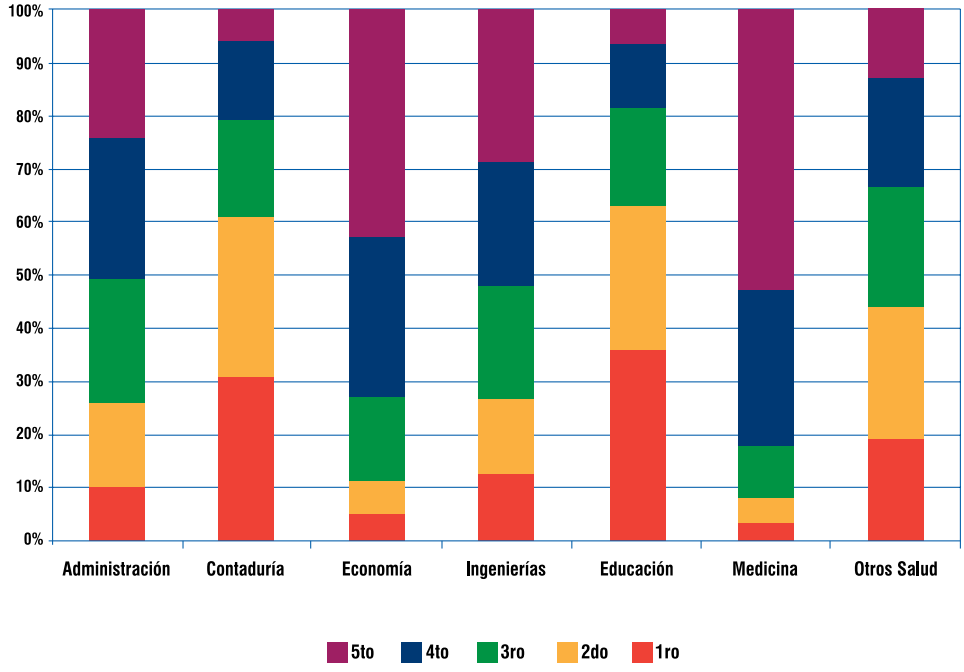
Tabla 4. Población evaluada en el piloto de la prueba GSA según programa e institución*

INSTITUCIÓN	Administración	Contaduría	Economía	Ingenierías	Educación	Medicina	Otros Salud	Total
Escuela Colombiana de Rehabilitación							84	84
Fundación Universidad del Norte	56		51	216	43	90	46	502
Fundación Universitaria Maria Cano	22	17					76	115
Institución Universitaria Colegios de Colombia							75	75
Politecnico Jaime Isaza Cadavid					141			141
Tecnológico de Antioquia					216			216
Uniminuto	28							28
Universidad de Antioquia					276			276
Universidad de la Salle	49	9	17	48	35		40	198
Universidad de Medellin	202	116	12	269				599
Universidad del Bosque	49					109	153	311
Universidad Industrial de Santander					25	43	29	97
Universidad Javeriana		11	80	175	158		69	493
Universidad Santo Tomás					17			17
Universidad Simón Bolivar		72		103			152	327
Total	**406**	**225**	**160**	**811**	**911**	**242**	**724**	**3479**

* El total reportado en esta tabla corresponde a 3.479 personas. La diferencia con respecto a los 3.757 evaluados mencionados anteriormente se debe a la exclusión de aquellos que no indicaron el semestre que cursaban y a quienes están matriculados en los semestres tercero, cuarto y quinto.

En cuanto a la distribución de la muestra evaluada por nivel socioeconómico, se observa que la población por quintiles del índice de nivel socioeconómico de los estudiantes (INSE)[4] varía sustancialmente de un área a otra. En el quintil de mayor nivel socioeconómico se concentran estudiantes de medicina y economía, un grupo importante de administradores e ingenieros y muy pocos estudiantes de contaduría, educación y otros programas de salud diferentes a medicina. Esta situación se invierte en el quintil más bajo **(Gráfico 2).**

Gráfico 2. Distribución de la población evaluada por quintiles del INSE según programas

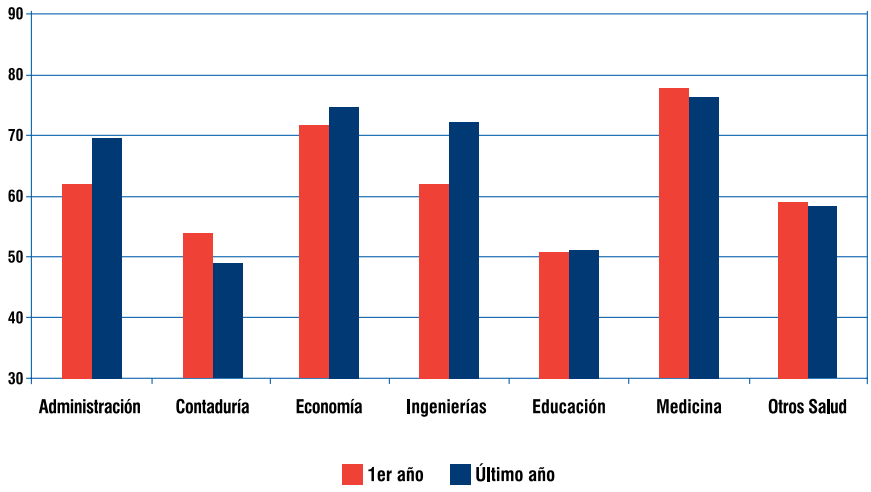


También se estimó el INSE promedio según el año cursado para cada uno de los programas, con el fin de detectar algún sesgo de tipo socioeconómico que pudiera distorsionar los resultados de la población evaluada. Se encontró que el INSE promedio es superior para los evaluados que cursan primer año de administración, economía, ingeniería y educación,

[4] El INSE es un indicador sintético construido por el ICFES mediante el empleo de técnicas estadísticas, como el análisis factorial y de componentes principales, a partir de la siguiente información aportada por los estudiantes en el momento de la inscripción: nivel educativo y ocupación del padre y de la madre; estrato; nivel en el SISBEN; personas por computador; conexión a internet en el hogar; ingresos familiares y hacinamiento (número de personas por habitación).

frente al INSE promedio de los que cursan el último año. La situación se invierte para los programas de contaduría y salud **(Gráfico 3).** De acuerdo con lo observado, las diferencias de desempeño no son atribuibles exclusivamente a las condiciones socioeconómicas de los estudiantes, sino principalmente a la experiencia universitaria.

Gráfico 3. Promedio del INSE según año cursado y programa



3. Resultados del pilotaje

A continuación se presentan los resultados de los análisis de los niveles de desempeño por género, año cursado, estrato socioeconómico, programa, área del conocimiento y dimensión cognitiva evaluada en la prueba. También se hace un análisis de la información psicométrica relevante obtenida de la aplicación, mencionando especialmente los resultados observados en los gráficos ítem-persona y los estadísticos de los ítems.

3.1 Niveles de desempeño

El desempeño por género fue muy similar en todas las dimensiones de la prueba, con excepción de solución de problemas, en la cual los hombres presentaron un mejor rendimiento que las mujeres. Los resultados promedio obtenidos son muy similares entre dimensiones evaluadas y entre géneros. La **Tabla 5** muestra los puntajes promedio y las desviaciones estándar reportados para hombres, mujeres y para el total de la muestra para cada dimensión cognitiva evaluada.

Tabla 5. Puntajes promedio y desviaciones estándar según género en cada dimensión cognitiva evaluada

GÉNERO	N	SOLUCIÓN DE PROBLEMAS		PENSAMIENTO CRÍTICO		ENTENDIMIENTO INTERPERSONAL		REPORTE		ARGUMENTO	
		Media	DE	Media	DE	Media	DE	Media	DE	Media	DE
Hombres	1.491	41,7	10,2	40,3	10,5	39,5	10,5	39,1	10,0	39,9	10,3
Mujeres	2.261	38,7	9,7	39,9	9,6	40,4	9,5	40,4	10,1	40,8	9,6
Ausentes	5	-	-	-	-	-	-	30,6	14,4	19,4	12,2
Total estudiantes	**3.757**	**39,9**	**10**	**40**	**10**	**40,1**	**9,9**	**39,8**	**10,1**	**40,2**	**10**

DE: desviación estándar

Con el fin de analizar si la prueba aplicada en Colombia mantiene su propiedad de validez discriminatoria, en la **Tabla 6** se contrastan los resultados obtenidos por los estudiantes de primer año con los de los alumnos de último año en cada una de las cinco dimensiones cognitivas. Se observó que los puntajes promedio de los estudiantes de último año son superiores a los logrados por los de primer año para todas las dimensiones evaluadas. Las mayores diferencias en los puntajes promedio se encontraron en pensamiento crítico, entendimiento interpersonal y redacción argumentativa.

Tabla 6. Puntajes promedio y desviaciones estándar según año cursado y dimensión cognitiva

AÑO	N	SOLUCIÓN DE PROBLEMAS		PENSAMIENTO CRÍTICO		ENTENDIMIENTO INTERPERSONAL		REPORTE		ARGUMENTO	
		Media	DE	Media	DE	Media	DE	Media	DE	Media	DE
Primero /1	1.683	38,1	9,5	37,7	9,6	37,9	9,8	38,6	9,8	38,2	9,4
Último /2	1.737	41,8	10,2	42,5	9,9	42,3	9,6	41,3	10,2	42,5	10,0
Total estudiantes /3	**3.757**	**39,9**	**10,0**	**40,0**	**10,0**	**40,1**	**9,9**	**39,8**	**10,1**	**40,2**	**10,0**

DE: desviación estándar

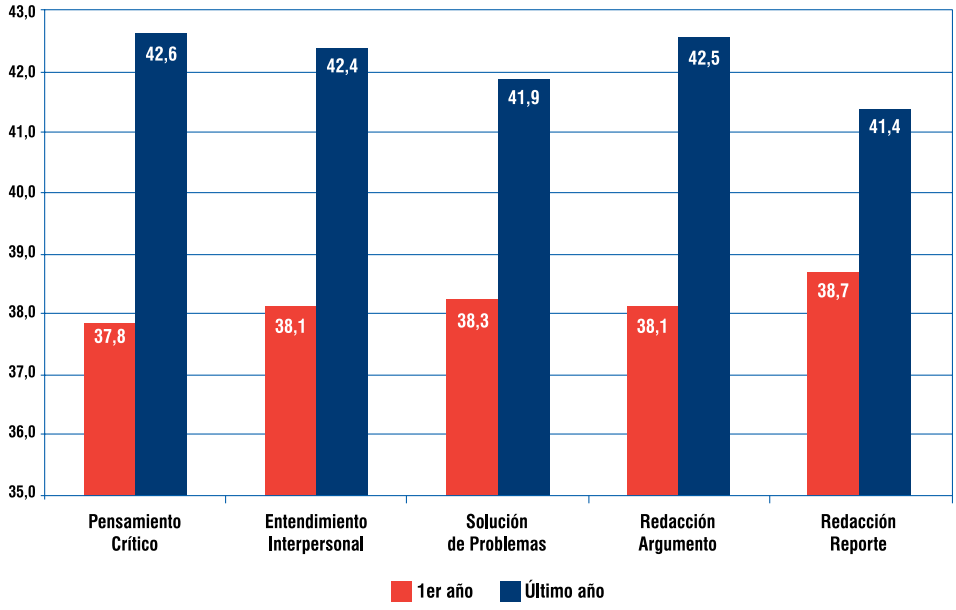
/1 Estudiantes de primer semestre.

/2 Estudiantes de semestres 6 - 13.

/3 Incluye estudiantes de 2º, 3º, 4º y 5º semestres (que suman 337 solamente, ya que no incluyen los de primer y último año) más los anteriores.

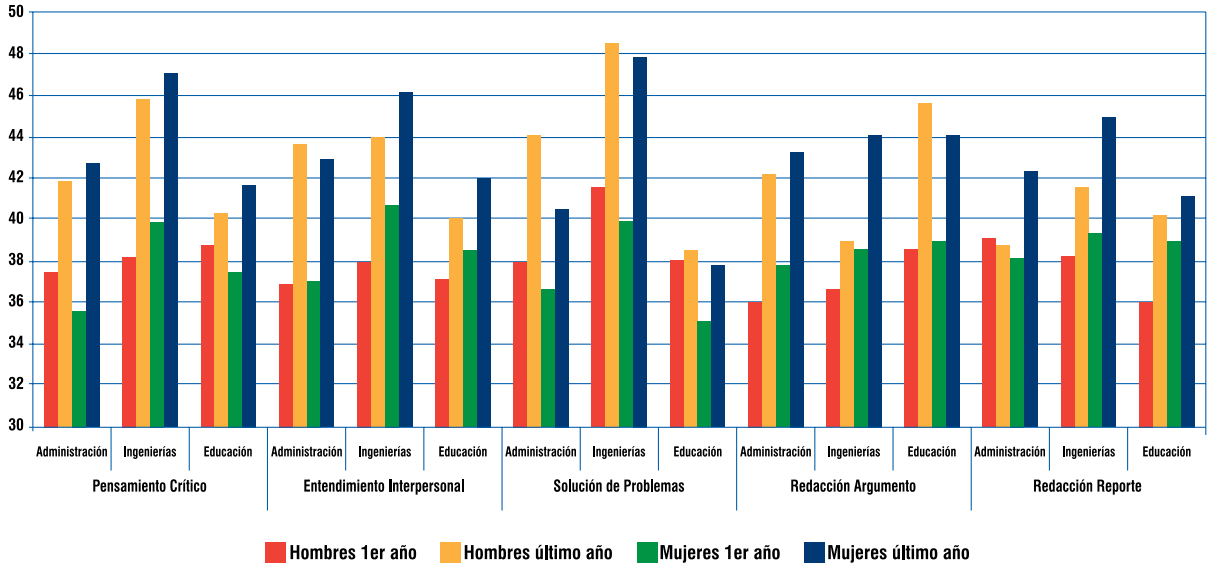
El **Gráfico 4** muestra la variación de los promedios presentados en la tabla anterior. Se observa un comportamiento similar en todos los casos, evidencia que permite inferir que la prueba GSA aplicada en el contexto colombiano mantiene su propiedad de validez discriminatoria. Esto quiere decir que el desempeño en la prueba está correlacionado positivamente con el éxito en la universidad y en el trabajo profesional, lo cual a su vez está vinculado con una amplia gama de conocimientos y habilidades curriculares que se adquieren mediante la experiencia universitaria.

Gráfico 4. Variación de puntajes promedio por año y dimensión cognitiva



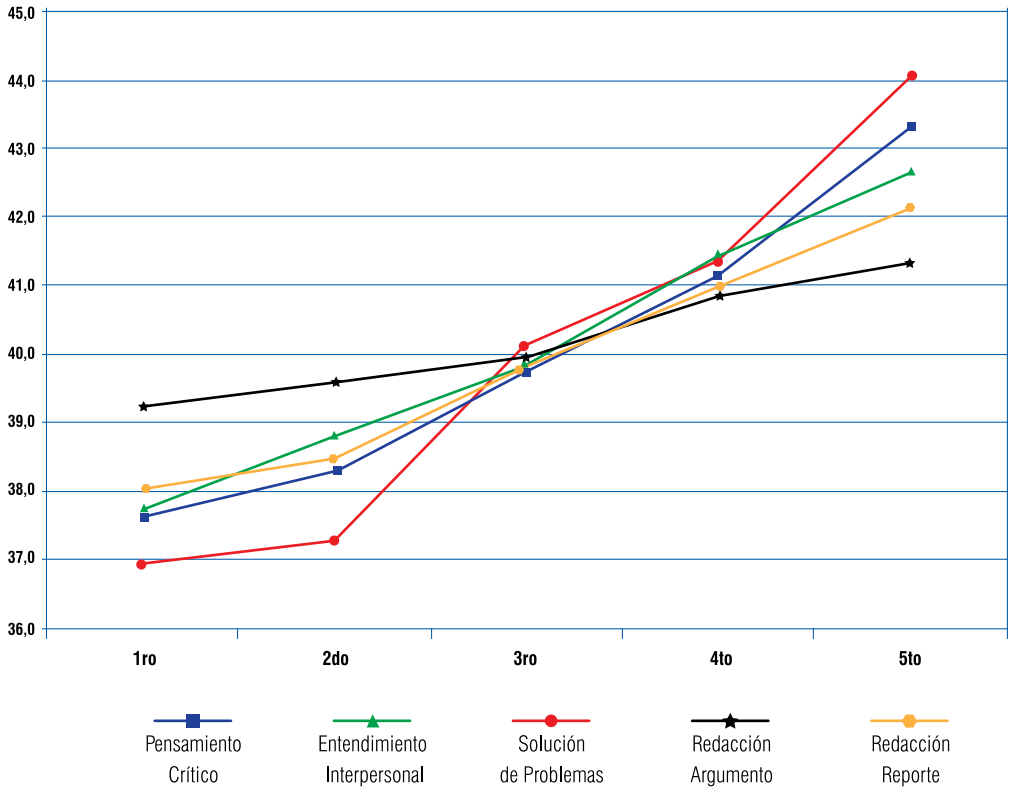
Al comparar las variaciones de puntajes promedio obtenidos por hombres y mujeres en el primer y último año, según programa cursado, se observa que las mayores diferencias en los puntajes según año cursado y, por ende los mayores niveles de logro, se concentran en el grupo de las mujeres **(Gráfico 5).**

Gráfico 5. Puntajes promedio según año cursado, género, programa y dimensión cognitiva evaluada



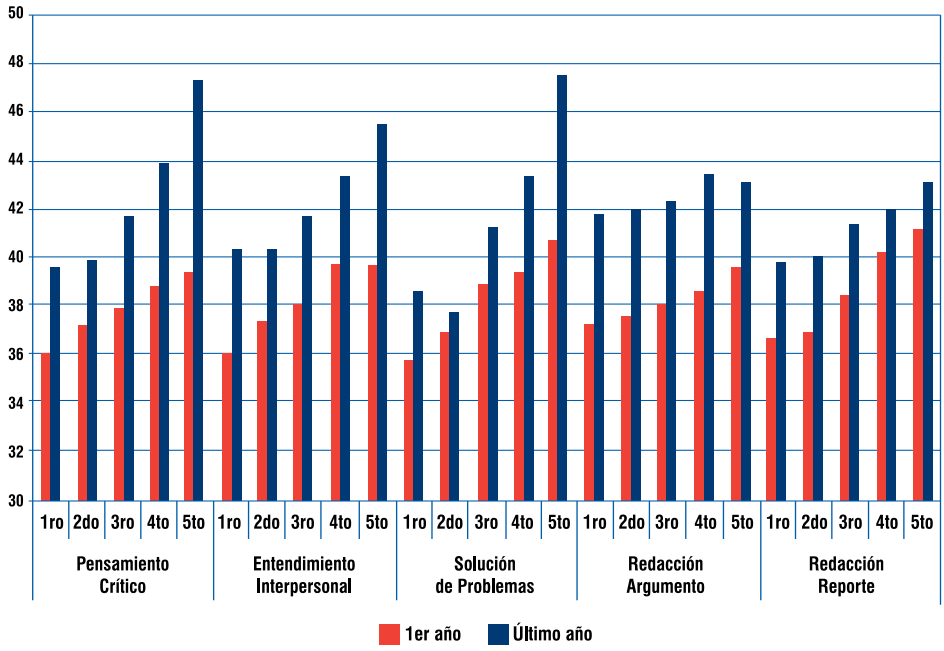
También se observó una relación directa entre el puntaje promedio logrado en cada dimensión cognitiva de la prueba y cada uno de los quintiles del INSE. En otras palabras, a mayor nivel socioeconómico, mayor puntaje promedio **(Gráfico 6).** La solución de problemas es la dimensión de la prueba que reportó las mayores variaciones, con más de 7 puntos promedio de diferencia entre los resultados obtenidos por la población del quintil más alto y la del quintil más bajo. Asimismo, la redacción argumentativa es la dimensión en la que se registran los menores contrastes en los puntajes promedio de los evaluados por quintiles.

Gráfico 6. Puntajes promedio según quintil del INSE y dimensión cognitiva



Al desagregar el resultado anterior por año cursado se halló una diferencia positiva en cada uno de los quintiles y dimensiones cognitivas, encontrándose las mayores variaciones en los puntajes promedio en el quintil más alto para todos los componentes, con excepción del de redacción **(Gráfico 7).**

Gráfico 7. Puntajes promedio según quintil, dimensión cognitiva y año cursado



En cuanto al desempeño por programas, es importante tener en cuenta que el número de estudiantes evaluados en cada programa varió significativamente, desde 779 en ingenierías a sólo 41 en el arquitectura / construcción de ambientes. En la **Tabla 7** se presentan los puntajes promedio y las desviaciones estándar para cada dimensión evaluada y por programas.

Tabla 7. Puntajes promedio y desviaciones estándar por dimensión cognitiva evaluada y programa

PROGRAMA	N	SOLUCIÓN DE PROBLEMAS		PENSAMIENTO CRÍTICO		ENTENDIMIENTO INTERPERSONAL		REPORTE		ARGUMENTO	
		Media	DE	Media	DE	Media	DE	Media	DE	Media	DE
Arquitectura / construcción de ambientes	41	39,7	8,9	38,4	8,5	39,5	10,1	36,4	10,9	42,5	11,3
Agricultura / reproducción animal / gestión ambiental	80	41,8	10,0	43,6	9,3	41,8	9,7	41,6	9,4	40,1	9,6
Negocios - finanzas	407	41,1	9,7	40,2	10,2	40,4	10,5	39,6	10,3	39,7	9,7
Negocios - administración / mercadeo	459	39,1	9,7	38,9	9,5	39,5	10,4	39,3	9,4	39,7	10,3
Educación - primaria / temprana edad	354	36,1	9,4	38,1	8,8	39,4	9,5	38,4	9,6	40,1	9,8
Educación - secundaria	610	37,6	9,8	40,4	9,9	39,9	10,1	40,3	11,1	42,8	11,1
Ingeniería	779	44,1	10,3	41,4	11,0	41,1	10,2	40,0	10,3	38,6	9,7
Salud - estudios médicos	439	41,4	9,7	41,8	10,6	40,8	9,7	41,9	9,7	40,4	8,6
Salud - enfermería	229	37,8	8,2	37,6	8,3	38,2	9,3	38,9	10,1	39,4	9,1
Salud - ciencias de la salud	354	37,4	8,4	38,4	8,6	39,2	8,4	39,4	9,0	40,4	9,4
Ausentes	5	-	-	-	-	-	-	30,6	14,4	19,4	12,2
Total estudiantes	**3.757**	**39,9**	**10,0**	**40,0**	**10,0**	**40,1**	**9,9**	**39,8**	**10,1**	**40,2**	**10,0**

DE: desviación estándar

Se observa que para la dimensión de solución de problemas el puntaje promedio más alto se presentó en ingenierías (44,1); para pensamiento crítico y entendimiento interpersonal esto sucedió en el programa de agricultura / reproducción animal / gestión ambiental (43,6 y 41,8, respectivamente); para comunicación escrita - reporte el promedio más alto se dio en el programa de salud – estudios médicos, para comunicación escrita – (41,9 puntos) y en argumento el puntaje más alta se observó en el programa de educación para el nivel de secundaria (42,8 puntos).

En contraste, los menores puntajes por programa en las dimensiones evaluadas fueron los siguientes:

- Solución de problemas: 36,1 en el programa de educación para el nivel de primaria y temprana edad
- Pensamiento crítico: 37,6 en el programa de salud - enfermería

- Entendimiento interpersonal: 38,2 en el programa de salud - enfermería
- Comunicación escrita – reporte: 36,4 en el programa de arquitectura / construcción de ambientes
- Comunicación escrita – argumento: 38,6 en el programa de ingenierías

En cuanto al desempeño por área del conocimiento, los resultados permiten observar que las que obtuvieron mayores puntajes promedio en los componentes de selección múltiple fueron arquitectura e ingeniería. Las mayores variaciones en los puntajes promedio también se presentaron en estas mismas áreas, con 43,9, y en ciencias de la educación, con 37,0; ambos en el componente de solución de problemas. Por su parte, en pensamiento crítico y entendimiento interpersonal no se presentaron grandes variaciones en los puntajes promedio para administración, contabilidad y economía, ciencias de la educación y ciencias de la salud. De igual forma, hubo poca diferencia en las variaciones promedio para todas las áreas del conocimiento en comunicación escrita - reporte, en la que se destaca ciencias de la salud con el mayor puntaje promedio, 40,3, y administración, contabilidad y economía con el menor, 39,3. El mayor puntaje promedio en comunicación escrita – argumento se presentó en ciencias de la educación, con 41,8, y el menor en arquitectura e ingeniería, con 38,7 **(Tabla 8).**

Tabla 8. Puntajes promedio y desviaciones estándar por área del conocimiento y dimensión cognitiva evaluada

ÁREA DEL CONOCIMIENTO	N	SOLUCIÓN DE PROBLEMAS		PENSAMIENTO CRÍTICO		ENTENDIMIENTO INTERPERSONAL		REPORTE		ARGUMENTO	
		Media	DE	Media	DE	Media	DE	Media	DE	Media	DE
Administración, contabilidad, economía, etc.	907	40,1	9,7	39,5	9,8	39,9	10,4	39,3	9,9	39,8	10,1
Arquitectura e ingenierías	859	43,9	10,3	41,6	10,8	41,2	10,2	40,1	10,2	38,7	9,7
Ciencias de la educación	964	37,0	9,7	39,5	9,6	39,7	9,9	39,6	10,6	41,8	10,7
Ciencias de la salud	1.022	39,2	9,1	39,7	9,6	39,7	9,2	40,3	9,6	40,2	9,0
Total estudiantes	**3.752**	**39,9**	**10,0**	**40,0**	**10,0**	**40,1**	**9,9**	**39,8**	**10,1**	**40,2**	**10,0**

DE: desviación estándar

Los resultados de desempeño se presentan por individuo y se reportan como logros por nivel para cada dimensión cognitiva evaluada. La **Tabla 9** muestra los puntajes requeridos para ubicarse en un logro en cada nivel según los componentes de la prueba GSA.

Tabla 9. Puntajes de individuos por niveles de desempeño

DIMENSIÓN COGNITIVA	DEBAJO DEL NIVEL 1	NIVEL 1	NIVEL 2	NIVEL 3 Y SUPERIOR
Solución de problemas	22 o menos	23 - 44	45 - 61	62 o más
Pensamiento crítico	24 o menos	25 - 44	45 - 61	62 o más
Entendimiento interpersonal	25 o menos	26 - 44	45 - 61	62 o más
Comunicación escrita - reporte	12 o menos	13 - 22	23 - 55	56 o más
Comunicación escrita - argumento	16 o menos	17 - 24	25 - 58	59 o más

La **Tabla 10** muestra el porcentaje de estudiantes que alcanzaron cada nivel de desempeño, en cada dimensión evaluada en la prueba. La alta concentración de la población en los niveles 1 y 2 demuestra que la GSA resultó muy difícil para el grupo objetivo. Este resultado puede ser producto de un factor desconocido del contexto, tal como la actitud de las personas al momento de responder la prueba. Por ello, se hace necesario realizar aplicaciones experimentales adicionales que permitan establecer un nivel apropiado de dificultad de los ítems.

Tabla 10. Distribución porcentual de personas según niveles de desempeño

DIMENSIÓN COGNITIVA	DEBAJO DEL NIVEL 1	NIVEL 1	NIVEL 2	NIVEL 3 Y SUPERIOR
Solución de problemas	5,5	65,3	27,7	1,5
Pensamiento crítico	7,6	61,3	29,5	1,6
Entendimiento interpersonal	9,0	62,6	26,6	1,8
Comunicación escrita - reporte	0,2	3,3	90,0	6,5
Comunicación escrita - argumento	1,9	3,3	92	2,8

Los **Gráficos 8** a **12** muestran los porcentajes de personas según niveles de desempeño en cada una de las dimensiones evaluadas, para cada programa y año cursado.

Gráfico 8. Distribución porcentual de personas según niveles de desempeño en pensamiento crítico según programa y año cursado

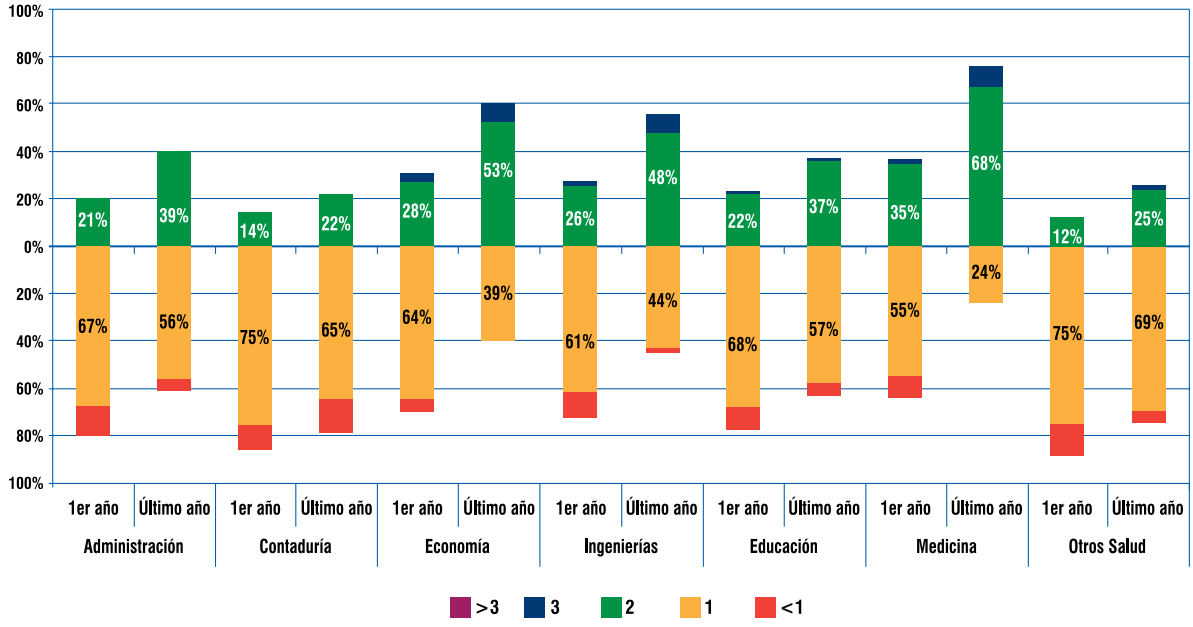


Gráfico 9. Distribución porcentual de personas según niveles de desempeño en entendimiento interpersonal según programa y año cursado

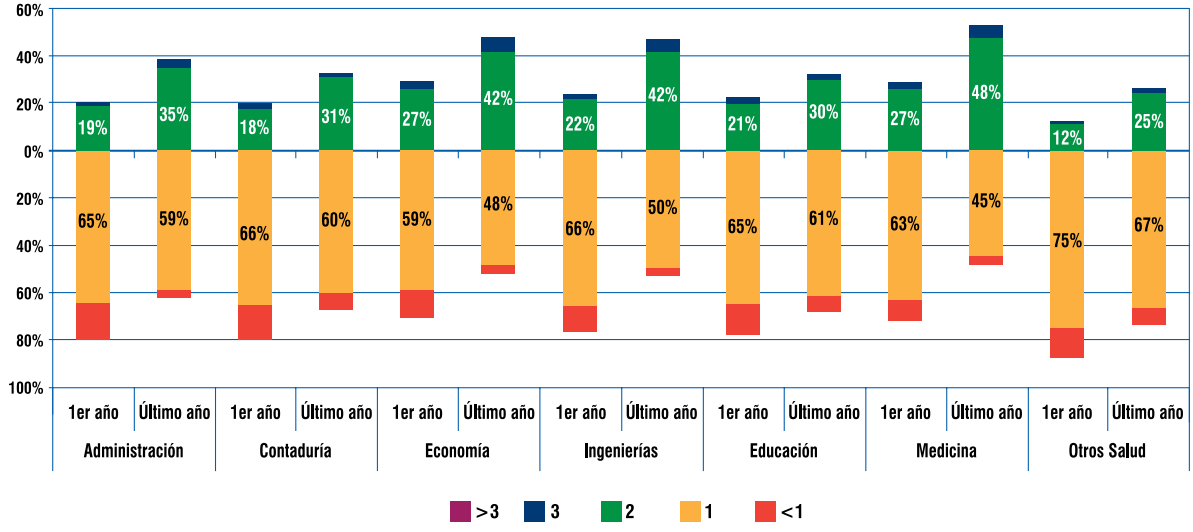


Gráfico 10. Distribución porcentual de personas según niveles de desempeño en solución de problemas según programa y año cursado

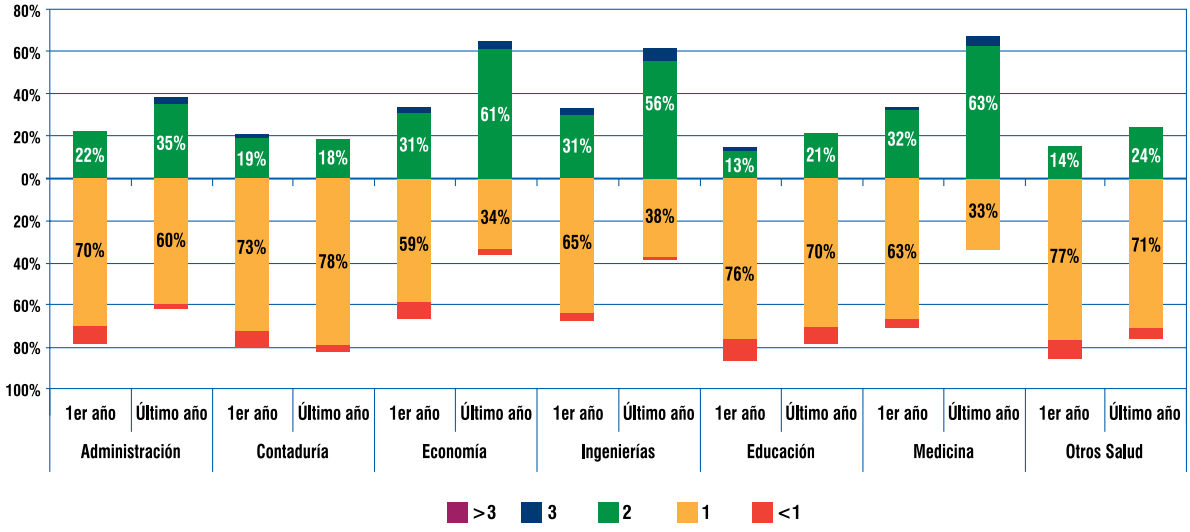


Gráfico 11. Distribución porcentual de personas según niveles de desempeño en comunicación escrita – argumento según programa y año cursado

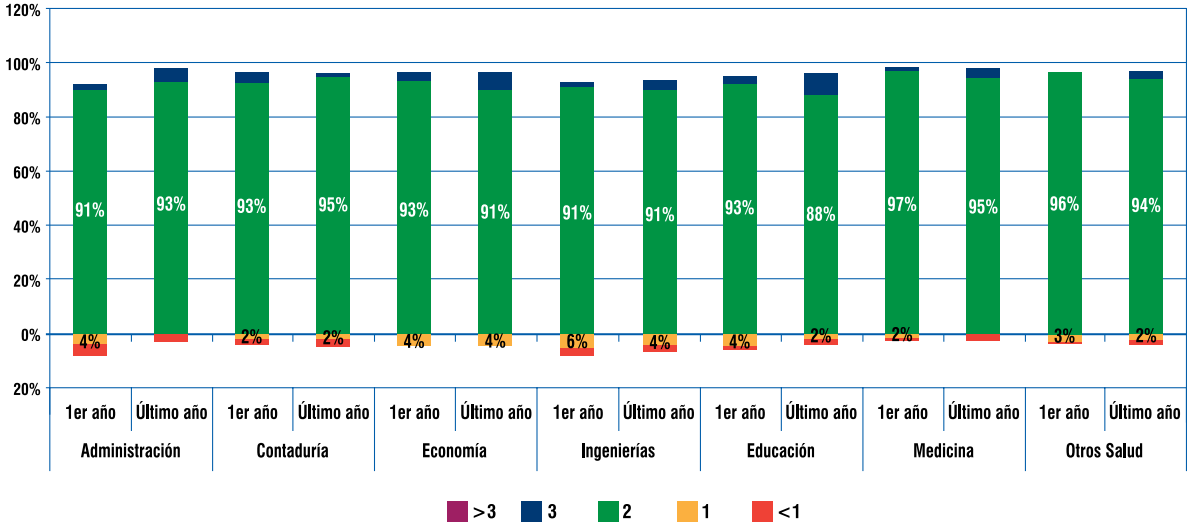
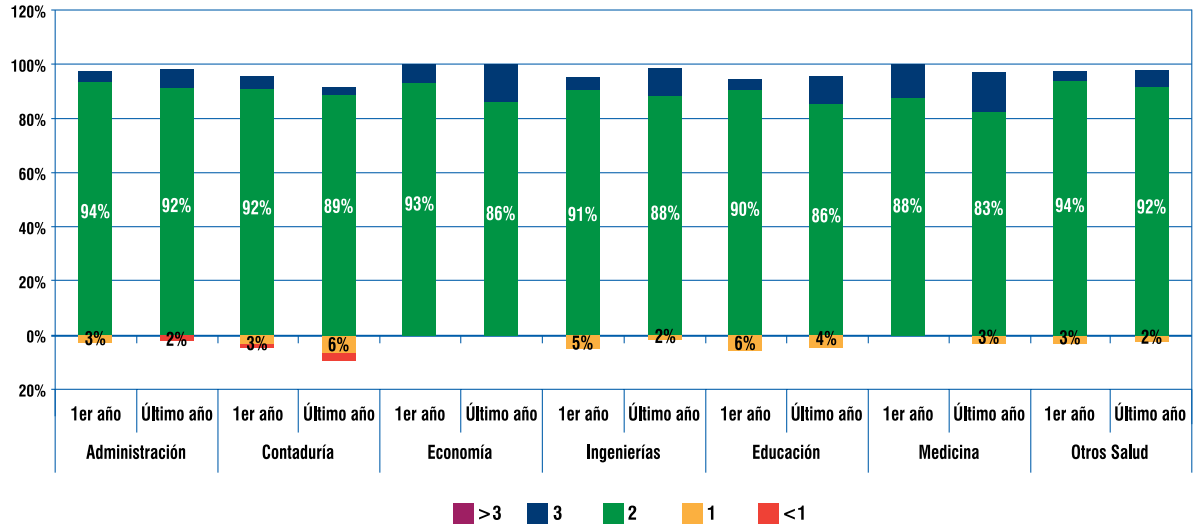


Gráfico 12. Distribución porcentual de personas según niveles de desempeño en comunicación escrita – reporte según programa y año cursado



A partir de la comparación del desempeño de los evaluados de primer y último año de los diferentes programas, el trabajo de Rosefsky y Saavedra (2011) investiga el valor agregado que las instituciones de educación superior brindan a sus estudiantes para el desarrollo de las habilidades genéricas. El estudio muestra el aumento de las competencias en pensamiento crítico, entendimiento interpersonal, solución de problemas y comunicación escrita en el transcurso de la formación universitaria, el cual es evidenciado con las disparidades en los resultados en la prueba GSA. De acuerdo con las contribuciones de la investigación, dicho valor agregado es proporcional a la relación cuantitativa de estudiantes por docente, y no tiene una correlación significativa con el porcentaje de profesores con doctorado en cada facultad, la dedicación del cuerpo docente, el proceso de admisión y el gasto por alumno en las universidades.

3.2 Análisis de los ítems

Se utilizaron técnicas de la Teoría de Respuesta al Ítem (TRI) en el análisis de los resultados (en los ítems de selección múltiple con doble respuesta se empleó el modelo de Rasch para estimar la habilidad del candidato y la dificultad del ítem). A continuación se presentan, para cada dimensión evaluada, gráficos ítem - persona y estadísticos de ítem del programa QUEST.

Los estadísticos correspondientes no se presentan para comunicación escrita porque el análisis output es complejo y no es fácilmente interpretable. Sin embargo, con un entrenamiento

continuo de marcadores y mediante la asignación de marcadores responsables de la asesoría en tareas múltiples será posible desarrollar análisis más sofisticados.

3.2.1 Gráficos ítem – persona

El gráfico ítem - persona es el resultado gráfico de QUEST. Este presenta tanto las estimaciones de dificultad del ítem y como las de habilidad del estudiante en la misma escala vertical. Se reportan las medidas en unidades llamadas logits. Se entiende que los estudiantes y los ítems situados en el extremo superior son más hábiles y más difíciles, respectivamente.

Cuando la habilidad de un estudiante se localiza en el mismo punto que una particular dificultad de ítem en dicha escala, el modelo de Rasch predice que el estudiante tendría una probabilidad del 50% de responder acertadamente el ítem. En este punto, el ítem proporciona la mayor cantidad de información posible sobre la habilidad de un estudiante y por lo tanto se maximiza la precisión de la medición de la habilidad que está siendo reportada.

Cuantos más ítems se sitúen en una región de la escala, más precisa será la medición de la habilidad en esta región. En caso de requerir alta discriminación en otra área de la escala, se debería considerar la utilización de un instrumento modificado.

En una escala de logit, la misma diferencia en logits siempre representa la misma diferencia en habilidad / dificultad.

Los **Gráficos 13** a **16** muestran los resultados ítem - persona para los componentes de selección múltiple de la prueba: solución de problemas, pensamiento crítico y entendimiento interpersonal. Por convención, el promedio de los valores logit para los ítems en el análisis se ubica en cero para definir el origen de la escala. Las personas están representadas por cruces a la izquierda de la escala vertical y los ítems por el número de ítem a la derecha de la escala. Como se mencionó, cuanto más alto esté situada una persona o un ítem en la escala, más hábil es la persona o más difícil es el ítem. La dispersión *(spread)* de los cruces da el rango de habilidad de las personas que contestan la prueba.

Los gráficos anteriores muestran que la prueba GSA fue, en general, muy difícil para las personas que la contestaron.

Gráfico 13. Ítem - persona para solución de problemas

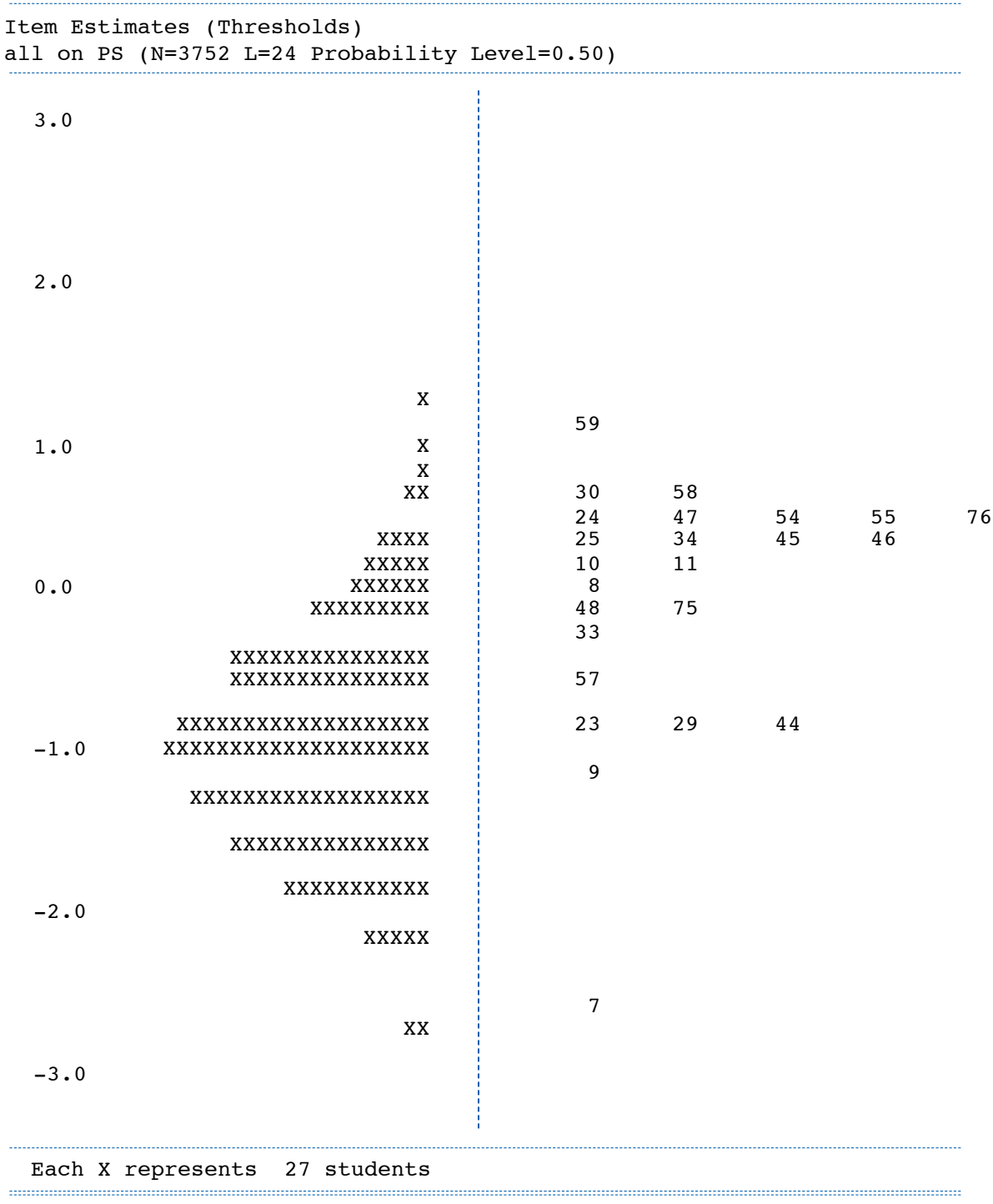


Gráfico 14. Ítem - persona para pensamiento crítico

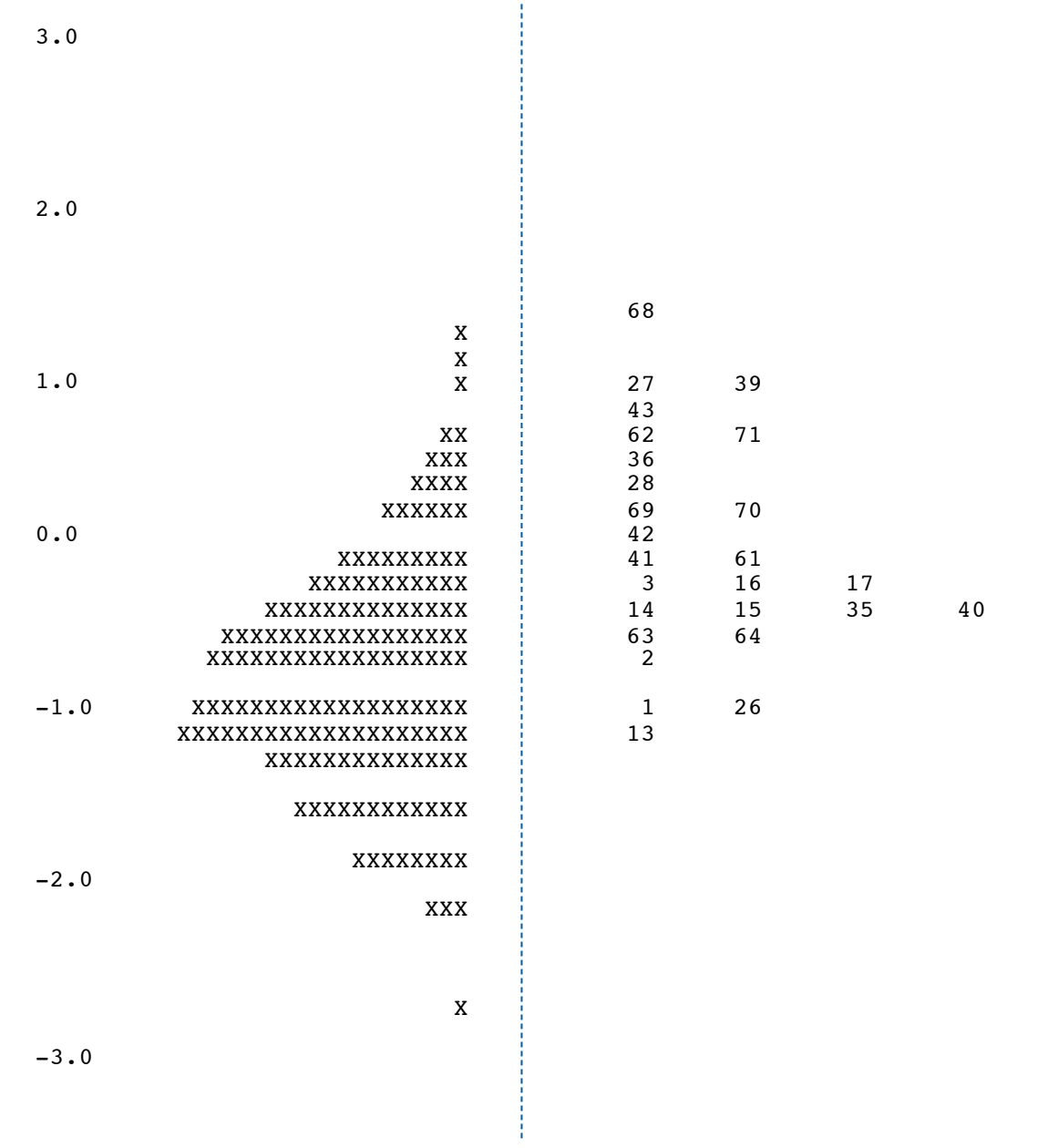
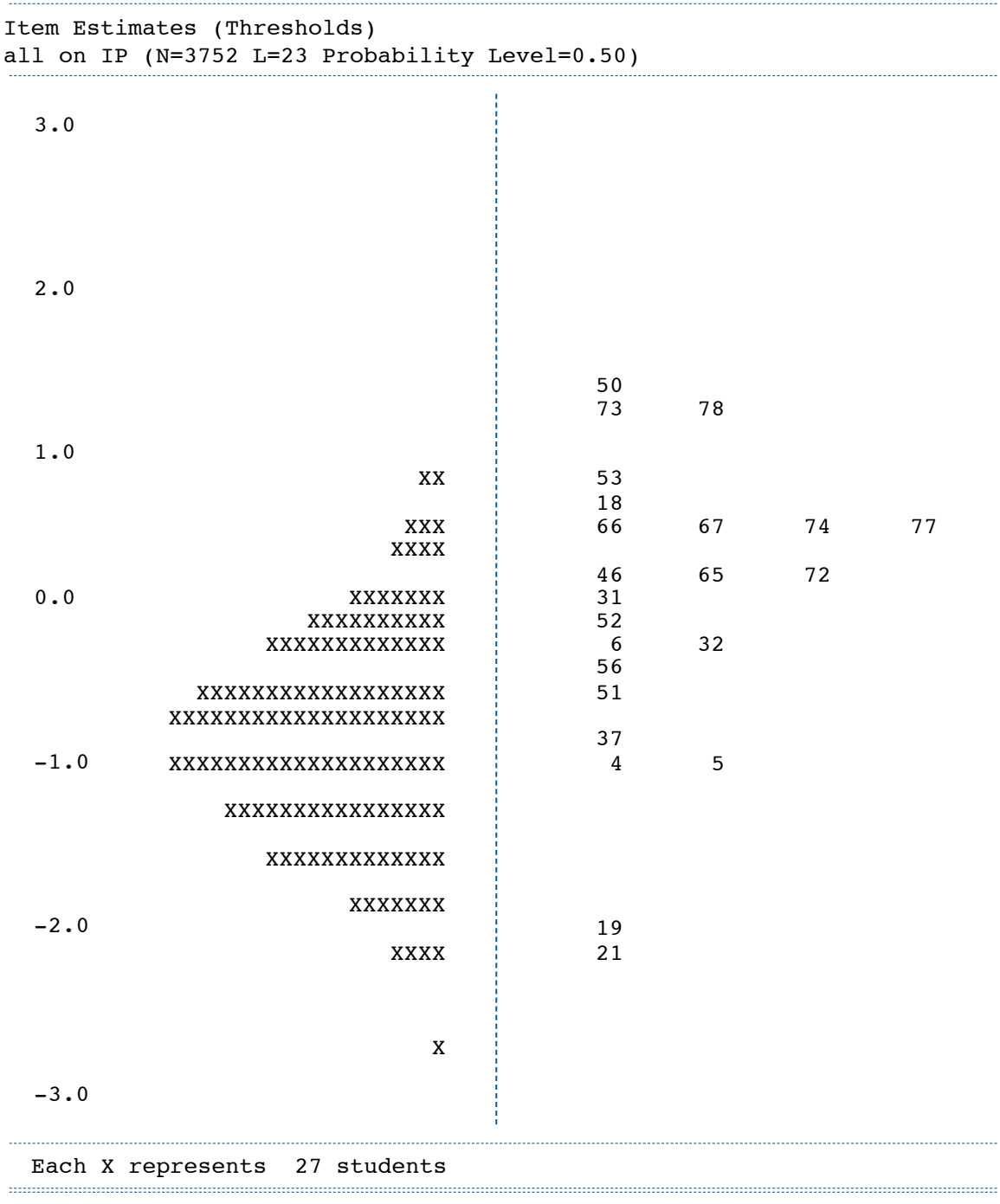


Gráfico 15. Ítem - persona para entendimiento interpersonal



3.2.2 Estadísticos de los ítems

Los estadísticos de ítems se presentan en las **Tablas 11**, **12** y **13**, las cuales corresponden a cada una de las tres dimensiones cognitivas consideradas y se dan para cada ítem numerado. Los estadísticos de ítem son los siguientes:

- **Facilidad:** es el porcentaje de estudiantes que responden correctamente.

- **Punto Biserial (Pt.Bis):** es un índice de la habilidad del ítem para discriminar entre estudiantes más y menos capaces (los ítems con punto biserial muy por debajo de 0,19 son raramente utilizados).

- **Diff:** es un índice de dificultad del ítem medido en logits (cuanto más alta sea la facilidad, menor el valor logit. Los logits son asignados de manera que el promedio del valor logit de los ítems para un análisis es igual a cero.

- **INFIT MNQS:** es un índice de qué tan bien se ajusta un ítem al desempeño de otros ítems que miden el constructo unidimensional (para medir qué tan bien se ajustan los ítems unos a otros con el fin de representar un único aspecto subyacente, se genera un estadístico de ajuste para cada pregunta. Este estadístico representa la diferencia entre la respuesta modelada y las respuestas observadas. El estadístico de ajuste medio cuadrado ponderado, *weighted mean square fit statistic*, INFIT MNQS, proporciona una medida de la coherencia del ítem al aspecto subyacente). De acuerdo con la Teoría de Respuesta al Ítem (TRI) de Rasch, aquellos ítems con valores INFIT MNQS relativamente altos deben ser examinados para ver si se desvían substancialmente de otros en el conjunto establecido para medir una dimensión singular y coherente.

- Adicionalmente, se proporciona un índice de confiabilidad de la prueba / componente, que en este caso es el Cronbach Alpha de Consistencia Interna, el cual se espera esté situado alrededor de 0,8.

Tabla 11. Estadísticos de ítems para la dimensión solución de problemas

Item	Facility	Pt Bis	p-value	Diff	Infit MNSQ
7	85.0	0.33	.000	-2.59	0.93
8	30.8	0.26	.000	0.12	1.04
9	58.8	0.37	.000	-1.13	0.96
10	28.0	0.39	.000	0.26	0.95
11	29.7	0.33	.000	0.17	0.99
23	49.8	0.36	.000	-0.74	0.97
24	23.3	0.35	.000	0.52	0.96
25	25.7	0.16	.000	0.39	1.10
29	50.0	0.43	.000	-0.75	0.92
30	21.6	0.30	.000	0.62	0.99
33	37.3	0.36	.000	-0.19	0.98
34	25.2	0.32	.000	0.41	0.99
44	50.2	0.41	.000	-0.76	0.93
45	25.8	0.20	.000	0.38	1.08
46	26.7	0.28	.000	0.32	1.02
47	22.9	0.33	.000	0.54	0.98
48	35.8	0.33	.000	-0.13	0.99
54	23.7	0.28	.000	0.49	1.01
55	24.5	0.24	.000	0.45	1.04
57	45.2	0.36	.000	-0.55	0.97
58	21.6	0.13	.000	0.62	1.11
59	14.4	0.17	.000	1.13	1.05
75	34.2	0.32	.000	-0.07	1.00
76	23.4	0.23	.000	0.50	1.04

Mean test score	8.01
Standard deviation	3.24

Tabla 12. Estadísticos de ítems para la dimensión de pensamiento crítico

Item	Facility	Pt Bis	p-value	Diff	Infit MNSQ
1	56.1	0.30	.000	-0.95	1.00
2	48.4	0.40	.000	-0.62	0.94
3	38.4	0.20	.000	-0.19	1.07
13	58.1	0.38	.000	-1.04	0.94
14	43.8	0.33	.000	-0.42	0.99
15	42.2	0.30	.000	-0.36	1.00
16	40.0	0.33	.000	-0.26	0.99
17	39.2	0.16	.000	-0.23	1.10
26	56.1	0.32	.000	-0.95	0.99
27	16.7	0.27	.000	1.00	0.98
28	27.2	0.27	.000	0.35	1.01
35	40.8	0.26	.000	-0.29	1.03
36	23.3	0.27	.000	0.57	1.00
39	17.8	0.30	.000	0.92	0.97
40	42.1	0.37	.000	-0.35	0.96
41	36.9	0.31	.000	-0.12	0.99
42	32.5	0.25	.000	0.09	1.03
43	19.5	0.11	.000	0.80	1.09
61	36.6	0.37	.000	-0.12	0.95
62	22.4	0.21	.000	0.61	1.04
63	44.7	0.44	.000	-0.48	0.91
64	44.7	0.33	.000	-0.47	0.98
68	11.6	0.16	.000	1.42	1.03
69	29.4	0.33	.000	0.22	0.97
70	30.2	0.28	.000	0.18	1.01
71	21.0	0.22	.000	0.69	1.02

Mean test score	9.05
Standard deviation	3.45

Tabla 13. Estadísticos de ítems para la dimensión de entendimiento interpersonal

Item*	Facility	Pt Bis	p-value	Diff	Infit MNSQ
4	55.5	0.26	.000	-0.91	1.02
5	57.2	0.25	.000	-0.98	1.03
6	39.6	0.22	.000	-0.23	1.05
18	21.1	0.04	.000	0.70	1.13
19	76.8	0.25	.000	-1.93	1.01
21	78.8	0.31	.000	-2.06	0.96
31	32.7	0.34	.000	0.08	0.96
32	39.8	0.30	.000	-0.24	1.00
37	51.7	0.37	.000	-0.75	0.96
49	28.5	0.23	.000	0.28	1.03
50	12.7	0.25	.000	1.31	0.98
51	45.6	0.38	.000	-0.49	0.95
52	35.4	0.37	.000	-0.05	0.95
53	19.3	0.22	.000	0.81	1.02
56	42.1	0.32	.000	-0.35	0.99
65	30.5	0.42	.000	0.17	0.91
66	24.4	0.30	.000	0.50	0.98
67	23.2	0.26	.000	0.56	1.00
72	30.4	0.31	.000	0.18	0.99
73	12.9	0.23	.000	1.28	0.99
74	25.1	0.21	.000	0.45	1.04
77	24.4	0.21	.000	0.49	1.04
78	14.1	0.23	.000	1.18	0.99

Mean test score	8.06
Standard deviation	2.79

* Los ítems 20, 22 y 38 fueron borrados de la dimensión de entendimiento interpersonal.

En los anteriores estadísticos de ítems se puede observar que la consistencia interna (Cronbach Alpha) para los 73 ítems de selección múltiple fue 0,75, lo cual es un nivel satisfactorio. La mayoría de ítems tuvo un valor de punto biserial superior a 0,2. Aquellos ítems con punto biserial bajo constituyen los de mayor dificultad para la población evaluada. No hubo ítems con valores INFIT MNQS particularmente altos (superiores a 1,20). Esto indica que los ítems no se desvían substancialmente del foco con respecto a los demás dentro del conjunto que busca medir una dimensión singular coherente.

4. Conclusiones

Los análisis de los hallazgos de la aplicación piloto de la prueba GSA en Colombia muestran que ésta fue demasiado difícil para los estudiantes seleccionados. Esto puede deberse a un factor desconocido: la actitud de los evaluados hacia la prueba. Los resultados en general mostraron que la dificultad de los ítems no coincidía adecuadamente con los rangos de habilidades de los estudiantes. Con el fin de tener un mejor entendimiento de la población estudiada se requiere mayor investigación por parte del ICFES y del ACER para establecer, por un lado, si lo observado es consecuencia del contexto en que los estudiantes presentaron la prueba y, por otro, un nivel apropiado de dificultad de los ítems.

Las estadísticas totales de la prueba fueron en general satisfactorias. Sin embargo, como el examen fue demasiado difícil para las personas, no era de esperarse una consistencia interna de los componentes.

Las correlaciones entre los puntajes de componentes de los estudiantes fueron en general apropiadas, de tal manera que mostraron suficiente *commonality* para poder usar los resultados totales de la prueba y la suficiente discriminación entre componentes para que los puntajes de los componentes individuales produzcan escalas significativas.

Lo anterior evidencia la necesidad de realizar nuevos estudios que posibiliten establecer niveles apropiados de dificultad de los ítems, lo cual permitirá hacer ajustes al instrumento que se aplicará en el país. Para esto es imperativo reunir más detalles sobre los estudiantes que participan en la prueba, tales como rendimiento anterior e indicadores de desempeño universitario actual. También es importante reunir mayores datos del contexto en el que se desempeñan los evaluados, lo cual debería permitir alinear las dificultades de los ítems a las habilidades de las personas.

5. Referencia bibliográfica

Rosefsky S., Anna; & Saavedra, J. E. (2011). “Do colleges cultivate critical thinking, problem solving, writing and interpersonal skills?”. En imprenta.

Calle 17 No. 3-40 • Teléfono:(57-1)338 7338 • Fax:(57-1)283 6778 • Bogotá - Colombia
www.icfes.gov.co